

University of Groningen

Analyzing longitudinal data with patients in different disease states during follow-up and death as final state

le Cessie, S.; de Vries, E.G.E.; Buijs, C.; Post, W.J.

Published in:
Statistics in Medicine

DOI:
[10.1002/sim.3755](https://doi.org/10.1002/sim.3755)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

le Cessie, S., de Vries, E. G. E., Buijs, C., & Post, W. J. (2009). Analyzing longitudinal data with patients in different disease states during follow-up and death as final state. *Statistics in Medicine*, 28(30), 3829-3843. <https://doi.org/10.1002/sim.3755>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Analyzing longitudinal data with patients in different disease states during follow-up and death as final state

Saskia le Cessie^{1,2,*}, Elisabeth G. E. de Vries³, Ciska Buijs³ and Wendy J. Post⁴

¹*Department of Medical Statistics and Bioinformatics, S5-P, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands*

²*Department of Clinical Epidemiology, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands*

³*Department of Medical Oncology, University Medical Centre Groningen, Groningen, The Netherlands*

⁴*Department of Epidemiology, University Medical Centre Groningen, Groningen, The Netherlands*

SUMMARY

This paper considers the analysis of longitudinal data complicated by the fact that during follow-up patients can be in different disease states, such as remission, relapse or death. If both the response of interest (for example, quality of life (QOL)) and the amount of missing data depend on this disease state, ignoring the disease state will yield biased means. Death as the final state is an additional complication because no measurements after death are taken and often the outcome of interest is undefined after death.

We discuss a new approach to model these types of data. In our approach the probability to be in each of the different disease states over time is estimated using multi-state models. In each different disease state, the conditional mean given the disease state is modeled directly. Generalized estimation equations are used to estimate the parameters of the conditional means, with inverse probability weights to account for unobserved responses.

This approach shows the effect of the disease state on the longitudinal response. Furthermore, it yields estimates of the overall mean response over time, either conditionally on being alive or after imputing predefined values for the response after death. Graphical methods to visualize the joint distribution of disease state and response are discussed.

As an example, the analysis of a Dutch randomized clinical trial for breast cancer is considered. In this study, the long-term impact on the QOL for two different chemotherapy schedules was studied with three disease states: alive without relapse, alive after relapse and death. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: longitudinal data; non-ignorable missing data; multi-state models; generalized estimation equations; inverse probability weighting; quality of life

*Correspondence to: Saskia le Cessie, Department of Medical Statistics and Bioinformatics, S5-P, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands.

†E-mail: cessie@lumc.nl

1. INTRODUCTION

The analysis of repeated measures is often hampered by missing observations. Patients can miss appointments, they do not always fill in questionnaires and patients can even drop out completely from the studies. This can be due to a variety of reasons, such as withdrawing consent to participate, lost to follow-up, disease relapse or death.

In this paper we consider the analysis of repeated measures with missing data of patients who can be in different disease states during follow-up, with death as the final state. Examples are cancer studies where patients can be disease free, alive after relapse or dead, or studies where patients with end stage renal failure can be on dialysis, alive after kidney transplantation or dead. The aim is to study the influence of the disease state on the longitudinal measurements and to estimate the marginal means over time.

There are several issues that need to be considered with these types of data. First, the disease state of the patient could be related both to the probability of missing measurements and to the longitudinal response measurements. For example, the quality of life (QOL) is generally lower after patients have a disease relapse when patients are also less likely to fill in QOL questionnaires. When the disease state is ignored, the probability of missing depends on the unobserved responses. This implies that the missing mechanism is not ignorable and that the disease state process should be taken into account when handling the missing observations. See [1] for a formal description of the different types of missing mechanisms. We assume in this paper that the probability of missing and the values of the missing observations may depend on the disease state and on the transition times of the disease state process. Within levels of the disease state and conditionally on the transition times, the data are assumed to be missing at random.

A second problem is missing observations due to death. Death differs from other reasons of missing where we can assume that the quantity of interest exists but is not measured. The definition of QOL after death is a topic of much discussion. Many statistical likelihood-based methods such as mixed effects models implicitly assume that the pattern of observation continues after death in a similar fashion as before death. This is often questionable. In the recent years dropout due to death has received considerable attention [2–7]. Death is frequently handled differently from other reasons of missing. Pauler *et al.* [2] use pattern mixture models with separate patterns for time of dropout and time of death. An alternative is to model the survival and response simultaneously. This is done by Ribaudo *et al.* [3], who uses random effect selection models and by Billingham and Abrams [4], who consider quality-adjusted survival analysis. Dufouil *et al.* [5] and Kurland and Heagerty [6] consider regression models conditioning on being alive. When applying a method it is important to keep in mind what the objective of a study is: the response which would have been observed if patients would not die versus the response conditional on being alive [8].

In this paper we combine the data on the disease state of a person with the response patterns over time. We use multi-state modeling to calculate the probability to be in a certain state over time. The mean response conditional on being in a certain disease state is estimated using generalized estimation equations (GEE) with an independent working correlation along the lines of Dufouil *et al.* [5] and Kurland and Heagerty [6]. Then the conditional means are averaged over the disease states to estimate the marginal response over time where death is treated as a special state. The conditional mean responses and the probabilities to be in each of the different disease states are also used to make inferences about the joint distribution of response and disease state.

We will apply the proposed methods to the data of a large multi-center clinical trial where two different chemotherapy schemes for patients with advanced breast cancer were compared [9, 10]. QOL was measured at various time points during follow-up. During the course of follow-up, three major disease states could be distinguished: disease free, alive after relapse and death.

2. THE DATA EXAMPLE

The data come from a large Dutch randomized multi-center study for patients with advanced breast cancer [9]. The aim of the study was to compare the long-term impact of two different chemotherapy treatments. A total of 885 patients were randomized and 838 patients were asked to participate in the health-related QOL (HRQOL) component of the trial [10]. Here, we restrict the analyses to the 838 participating patients of which 422 patients received a standard dose of chemotherapy and 416 patients received a high dose of chemotherapy. While only 804 patients actually participated in the HRQOL study, the data of survival and disease status were available for all 838 patients.

During follow-up, patients could be in the three major disease states: alive and disease free, alive after relapse and death. Figure 1 shows a graphical representation of the multi-state situation for these data.

The SF 36 questionnaire was used to measure the different aspects of QOL. In this paper we use the subdomain physical health (PH) to illustrate our proposed methodology. PH was measured at start treatment (visit 0), at 3 months (just after chemotherapy stopped), at 6 months after start systemic treatment and thereafter every half year. We consider here measurements until 5 years after follow-up. The percentage of missing observations increased over time from 10 per cent at the first visit to more than 60 per cent at the last visit. Missing was related to the condition of a patient. Obviously, there are no observations after death. Furthermore, after a relapse of disease the percentage of missing responses increased because patients were less inclined to fill in the questionnaires after disease progression. Moreover, some centers did not hand out QOL questionnaires to patients after a relapse occurred. The average percentage of missing values for patients alive after a relapse was around 40 per cent.

The high-dose chemotherapy tends to have a positive effect on relapse-free survival [9], but it has a large negative impact on the PH in the first year (visit 1–3) [10]. The question addressed in this paper is whether the high-dose chemotherapy also has a long-term effect on the PH or

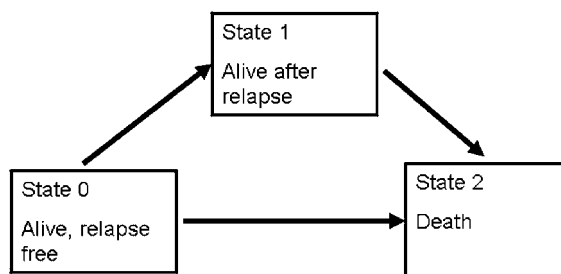


Figure 1. Graphical representation of the disease states of the breast cancer data.

whether the negative effects disappear after the first year of follow-up. Furthermore, we will study the effect of disease stage on the PH.

3. NOTATIONS AND GENERAL APPROACH

We start by defining some notations. There are I subjects and the longitudinal outcome Y_{ij} is scheduled to be measured at J time points t_1, \dots, t_J , with Y_{ij} as the measurement of subject i on time point t_j . We work in the generalized linear models context and the repeated measures can be either binary, count data or continuous. Subjects can be in a finite number of different disease states $s = 0, \dots, S$ and S_{ij} indicates the disease state of subject i at time t_j . All patients start at $t = 0$ in state 0. The vector of baseline covariates for person i is X_i , whose value does not change during the duration of the study. In the data example, we consider one binary covariate X for the treatment group.

The idea is to model the mean response conditionally on being in a certain state s : $\mu_{ij}^s = E[Y_{ij}|X_i, S_{ij}=s]$. The probability to be in state s at time t , $P_s(t|X_i)$, is obtained using multi-state models. Subsequently, the marginal response over time, treating death as a special situation, and the joint distribution of response and state are estimated.

4. ESTIMATING DISEASE STATE PROBABILITIES USING MULTI-STATE MODELS

The probability to be in disease state s at time t , given the covariates X , $P_s(t|X)$ can be estimated using Markov multi-state models. Multi-state models consider the situation where different events can occur over time and patients can be in different disease states. In a Markov multi-state model, the transition rates to new states, given the event history, only depend on the present state. Putter *et al.* [11] wrote a tutorial on competing risk and multi-state models and details about definitions and estimation issues can be found therein. In the remainder of the paper, we focus on the simple illness–death model in detail as depicted in Figure 1, with a disease-free state (state 0), a state where patients are ill, but still alive (state 1) and death (state 2), but results can be extended to more complex multi-state models. We denote the hazard rate (transition intensity) given the vector of baseline covariates X of the transition from state i to j by $h_{ij}(t|X)$.

In the illness–death model, the disease state process is completely characterized by the disease-free time T^1 , the time between start and illness, and by the survival time T^2 , the time between start and reaching state 2. Both T^1 and T^2 can be censored, for example if a patient dies while being disease free or if a patient is alive at the end of follow-up.

If the covariate vector X consists of only a few categorical variables, the transition hazards $h_{01}(t|X)$, $h_{02}(t|X)$ and $h_{12}(t|X)$ can be estimated using separate Nelson–Aalen estimators for each covariate pattern. Otherwise these transition hazards can be estimated using the Cox proportional hazards model assuming the covariates to act multiplicatively on the baseline hazard. In case of many states and many transition hazards more assumptions can be made such as assuming that different transition hazards are proportional but we will not discuss this further here. There are two options for the time scale in state 1, the clock forward approach, using time from diagnosis as time scale, or the clock reset approach, where the clock is reset to 0 in state 1. The probabilities to be in each of the different states at a certain time point can be derived from the transition hazards.

To remain in state 0 at a certain time point, a patient should not have transferred to state 1 or state 2 before this time. The hazard to transfer to either state 1 or 2 at time u is $h_{01}(u|X) + h_{02}(u|X)$. The probability to be in state 0 at time t is therefore

$$P_0(t|X) = e^{-H_{01}(t|X) - H_{02}(t|X)}$$

with $H_{ij}(t|X) = \int_0^t h_{ij}(u|X) du$, the cumulative transition hazard to transfer from state i to state j .

To be at state 1 at time t , a patient should have stayed until a certain time point $u < t$ in state 0, transferred from state 0 to state 1 at time u , and remained in state 1 (hence not move to state 2) between u and t . In case of a clock forward approach, the cumulative hazard to move from state 1 to state 2 between time u and t is $H_{12}(t|X) - H_{12}(u|X)$ and the probability to remain in state 1 between time u and time t is therefore $\exp(-H_{12}(t|X) + H_{12}(u|X))$. Therefore, the probability to be in state 1 at time t is obtained by integrating over u :

$$P_1(t|X) = \int_0^t e^{-H_{01}(u|X) - H_{02}(u|X)} h_{01}(u) e^{-H_{12}(t|X) + H_{12}(u|X)} du$$

Finally, $P_2(t|X) = 1 - P_1(t|X) - P_0(t|X)$. When using the clock reset approach the formula for $P_1(t|X)$ has to be adapted. Details on this can be found in [11].

5. MODELING RESPONSES PER DISEASE STATE

5.1. Modeling conditional means over time

This section describes how μ_{ij}^s , the mean response of subject i at time t_j , conditionally on being in state s , can be estimated. Mixed effects models are commonly used to model repeated measures over time. These are the so-called subject-specific models, where each subject has its own individual pattern over time and random effects are used to model the intersubject variability. These models are not useful here. If a subject drops out, implicitly its trajectory is extrapolated and the missing values are imputed. This results in an estimated mean pattern over time, which reflects the effects that would have been observed if patients never drop out of the state. For example, if the individual mean response is constant over time in a state, but subjects with a lower response are more likely to transfer to another state, then the marginal mean conditional on being in the state will increase. Mixed effects models estimate the individual response over time and will estimate a constant effect which is in this situation too low. See [5, 6, 8] for more discussion on the different aims in modeling repeated measurements.

We are interested in the mean response pattern over time, conditional on being in a certain state. Therefore, we use the results of Kurland and Heagerty [6] who study regression models conditioning on being alive. Here marginal models, also termed as population-averaged models, are used to model the conditional mean [12]. They show that the conditional mean can be consistently estimated using GEE, as long as an independent working correlation structure is used. This can be used analogously if the conditional mean is modeled separately per state. The conditional mean μ_{ij}^S is modeled directly by a (generalized) linear model, for example a linear regression model in case of continuous outcomes or a logistic model in case of binary outcomes. The GEE equations

for the conditional regression model are

$$U(\beta^S) = \sum_{i=1}^I \sum_{j=1}^J \frac{\partial \mu_{ij}^S}{\partial \beta^S} \sigma_{ij}^{-2} I_{[S_{ij}=s]} (Y_{ij} - \mu_{ij}^S)$$

with $\sigma_{ij}^2 = \text{var}(Y_{ij})$ and $I_{[x]}$ is the indicator function, being equal to 1 if x is true and 0 otherwise. The GEE estimates are the solutions to $U(\beta^S) = 0$. If the regression model is correctly specified, solving the GEE approach yields under standard regularity conditions consistent estimates of the parameters and their standard errors.

5.2. Handling missing observations

We consider different types of missing observations: missing observations while the disease state of a patient is still observed or missing observations due to incomplete follow-up of a patient. In the last situation, the patient drops out of the study alive before the last scheduled measurement and the actual disease state is unknown after the time of dropout. The event times T^2 and probably T^1 , if the dropout is in the disease-free state, are censored. Furthermore after death, the response is always missing.

To handle the different types of missing values, two indicator variables are defined. The dropout indicator variable C_{ij} is equal to 1 if the measurement of subject i at time point j was taken before dropout and 0 otherwise. The second missing indicator variable R_{ij} indicates whether Y_{ij} is observed ($R_{ij} = 1$) or not ($R_{ij} = 0$). This missing response process R is censored at the time of dropout or death.

We now consider GEE equations where summation is done over the observations actually observed (hence, both $R_{ij} = 1$ and $C_{ij} = 1$) within a state. This yields as actually computed GEE equations for estimating the conditional mean in state s :

$$U_{\text{comp}}(\beta^S) = \sum_{i=1}^I \sum_{j=1}^J \frac{\partial \mu_{ij}^S}{\partial \beta^S} \sigma_{ij}^{-2} I_{[S_{ij}=s]} C_{ij} R_{ij} (Y_{ij} - \mu_{ij}^S) = 0 \quad (1)$$

If $E_{R,C,Y|S}[U_{\text{comp}}(\beta^S)] = 0$, the estimates obtained by solving these GEE equations are consistent. We now discuss conditions for R and C for which this is true and how the GEE equations can be modified if this does not hold.

Let Y_i be the complete vector of (observed and unobserved) responses of subject i . The condition $E_{R,C,Y|S}[U_{\text{comp}}(\beta^S)] = 0$ is satisfied if $E_{R,Y|S}[R_{ij}(Y_{ij} - \mu_{ij}^S)] = 0$, under the assumption that the dropout process is independent of the actual survival time or disease-free time, and does not depend on the past or future responses of a subject. Hence, we assume that:

$$P(C_{ij} = 1 | T_i^1, T_i^2, Y_i) = P(C_{ij} = 1)$$

The assumption of independent dropout is a rather strict assumption but not unreasonable in situations where the major reason for censoring of event times is administrative.

Kurland and Heagerty [6] distinguished different types of missing response processes R , depending on how the probability $P(R_{ij} = 1 | T^1, T^2, Y)$ can be reduced. They show that $E_{R,Y|S}[R_{ij}(Y_{ij} - \mu_{ij}^S)] = 0$ and that solving the computed GEE equations (1) yield consistent estimates, if the missing response process neither depends on the response history nor on the disease state process (missing completely at random (MCAR)).

However, disease state and the past responses often influence the probability of responding at a certain time point. Therefore, we consider a less restricted process for R , in the line of MAR-S of Kurland and Heagerty [6], who defined a missing at random process conditional on the exact survival time.

We assume that the Y_{ij} are monotone missing, meaning that if a person does not respond at a certain time point, he will not respond at subsequent time points, and assume that all subjects respond at the first measurement. Furthermore, we assume that the probability of observing a response may depend on the previous observed responses, and on the disease free and survival time, but not on the current or on future values of the response. This yields the following mechanism for the missing responses

$$P(R_{ij}=1|T_i^1, T_i^2, Y_i) = P(R_{ij}=1|\bar{Y}_{ij}, T_i^1, T_i^2)$$

with \bar{Y}_{ij} being the subject's response history, i.e. the measurements of person i until time point t_j but not including Y_{ij} . Conditional on disease-free time and survival time, this is a missing at random mechanism. The situation that the response probability depends on the actual survival or disease-free time could happen, for example if patients just before the occurrence of a relapse are less inclined to fill in QOL questionnaires.

Inverse probability weighted GEE [13] is a way to deal with missing at random responses. We consider the following equations:

$$U_{\text{weight}}(\beta^S) = \sum_{i=1}^I \sum_{j=1}^J \frac{\partial \mu_{ij}^S}{\partial \beta^S} \sigma_{ij}^{-2} I_{[S_{ij}=s]} w_{ij} R_{ij} C_{ij} (Y_{ij} - \mu_{ij}^S) \quad (2)$$

If μ_{ij}^S and R_{ij} are modeled correctly, and weights are chosen such that $w_{ij} = 1/P(R_{ij}=1|\bar{Y}_{ij}, T_i^1, T_i^2, S_{ij}=s)$, solving these weighted equations still yields consistent estimates, because the expectation of (2) is then still equal to 0. Note that $P(R_{ij}=1|\bar{Y}_{ij}, T_i^1, T_i^2, S_{ij}=s) = P(R_{ij}=1|\bar{Y}_{ij}, T_i^1, T_i^2)$, because the disease status process is completely determined by T_i^1 and T_i^2 . The weights for monotone dropout can be estimated using life table methods (see [5, 6]). A response is observed at time point t_j if a subject responded at all previous time points, and did not stop responding at t_j . This means that

$$P(R_{ij}=1|\bar{Y}_{ij}, T_i^1, T_i^2) = \prod_{k=2}^j (1 - \pi_{ik})$$

with $\pi_{ik} = P(R_{ik}=0|R_{i(k-1)}=1, \bar{Y}_{ik}, T_i^1, T_i^2)$ being the probability for subject i to stop responding at time point t_k . The missing response process is censored after death or at the end of follow-up. The probabilities π_{ik} can be estimated using logistic regression, with previous responses, the disease free and survival time, and other covariates as independent variables.

An additional problem is that the disease free and survival time are not known for all subjects. In the case of censored event times at the end of the intended follow-up, one could use the information that $T^2 > t_J$ or $T^1 > t_J$ in the missing response model instead of the exact event times, assuming that this captures sufficient information.

If there are also subjects with earlier censored event times, different approaches can be followed. A simple, but inefficient solution is to leave out subjects who could not be followed for the complete intended follow-up period. Imputing the missing survival and relapse times could be a

more efficient alternative. A third approach is to make additional assumptions about the missing response model. For example in certain situations, the exact survival time in the missing response model could be replaced by a dichotomous variable, which indicates if a person dies before the next visit. In the breast cancer trial example we followed this approach.

Standard errors for the regression coefficients can be calculated using the sandwich estimator, but this does not account for the fact that the weights are estimated. Therefore, we recommend using bootstrapping. This is discussed further in the next section.

6. COMBINING CONDITIONAL RESPONSES PER STATE AND DISEASE STATE PROBABILITIES

The marginal mean response over time can be obtained from the disease state probabilities and the conditional means per state, using that: $E[Y_{ij}|X_i] = \sum_s P_s(t_j|X_i)\mu_{ij}^S$. As the disease state probabilities are not linear in X , the marginal mean $E[Y|X]$ is in general not linear in X . For a binary exposure variable, p -values and confidence intervals for the marginal effect of X can be obtained by bootstrapping. For continuous covariates the relation between $E[Y|X]$ and X can be visualized by plotting the marginal means over time for several levels of X .

A special case here is the state death for which no response values are available. Predefined values could be imputed here, such as for example a value of 0 for QOL after death. However, it is often argued that it is meaningless to consider QOL after death and that QOL should only be considered on the conditioning of being alive. In this case the marginal mean is:

$$E[Y_{ij}|X_i, S_{ij} \neq 2] = \sum_{s=0}^1 P_s(t_j|X_i)\mu_{ij}^S / (1 - P_2(t_j|X_i))$$

To calculate the standard errors of the marginal means, both the error in the conditional means and in the multi-state probabilities have to be taken into account. It is complicated to calculate the standard errors analytically. The standard errors of the conditional means could be calculated using the sandwich estimator, but this ignores that the probability weights are estimated. Therefore, we suggest using bootstrapping to calculate standard errors. The resampling should be done from individuals rather than from observations, to preserve the correlation between observations of the same subject [14] and in each bootstrap data set, the multi-state probabilities and the weights for the GEE should be recalculated.

In clinical practice, the joint distribution of disease state and response over time is of major interest. This joint distribution can be estimated straightforwardly from the conditional response over time given the disease state and the disease state probabilities, using that $f(Y, S|t) = f(Y|S, t)f(S|t)$. If Y is binary, the distribution of Y given S follows directly from the estimated mean given S . If Y is continuous, assumptions about the distribution of Y have to be made, such as assuming that Y is normally distributed. In such a situation it is often more illustrative to select a cut-off value below which a response is defined as low, and calculate $f(Y < c, S|t)$. The joint probabilities obtained in this way could be used to make partitioned survival plots, similar to the Q-TWiST plots described in Glaziou *et al.* [15].

7. SIMULATION

We performed a simulation study to demonstrate the value of accounting for the disease state. We considered three disease states: disease free (state 0), illness (state 1) and death (state 2). In states 0 and 1 repeated outcome responses were simulated for 400 subjects at 11 visits ($t_j = 0, 1, \dots, 10$) from the model $Y_{ij} = \alpha_i + 80 \times I_{[s_{ij}=0]} + 50 \times I_{[s_{ij}=1]} + \varepsilon_{ij}$. Here α_i , the random subject effect, followed a normal (0, 100) distribution and ε_{ij} a normal (0, 4) distribution. The disease-free time T^1 was simulated from an exponential $(1.3 - 0.6I_{[\alpha_i < 0]})$ distribution, and the survival time T^2 from an exponential $(0.65 - 0.3I_{[\alpha_i < 0]})$ distribution, such that subjects with a negative value for α_i had a higher probability on relapse or death.

In this complete data set, missing observations were simulated. The probability of a first missing response at t_j , π_{ij} was equal to 0.05 in the disease-free state and 0.10 in the disease state. The time of dropout was drawn from an exponential (0.08) distribution, survival and relapse time were censored at this time and the response Y was missing after the dropout time.

Three different approaches were applied on the data set with missing observations. We applied the method described in this paper, accounting for and averaging over the disease state, using GEE and multi-state probabilities. The second method was GEE conditional on being alive, ignoring the disease state, along the lines of [6]. The third approach accounted for disease state, but used linear mixed models with a random person effect to estimate the mean response per state. The estimated mean responses of the three approaches were compared with the observed marginal responses conditional on being alive in the complete data set. Figure 2 shows the average results of 1000 simulation runs.

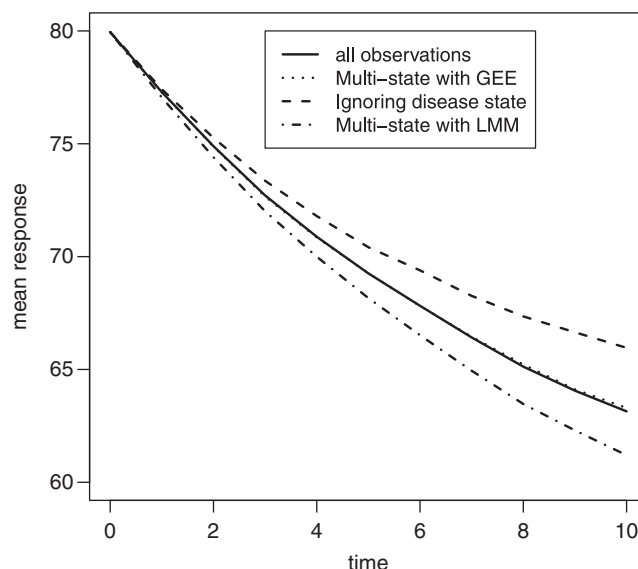


Figure 2. Comparison of three approaches to estimate the marginal mean conditional on being alive: accounting for and averaging over disease state using GEE (multi state with GEE), using GEE but ignoring the disease state (ignoring disease state) and accounting for disease state and averaging using linear mixed models (multi state with LMM). Estimated means are compared with the observed marginal means in the complete data set. The results given are the averages of 1000 simulation runs.

From this figure it is clear that our approach of accounting for and averaging over disease states yields unbiased estimates of the marginal mean over time, conditional on being alive, while the other two approaches are biased. Ignoring the disease states leads to estimated means that are too high because the frequency of missing observations is larger in the disease state. The use of mixed models per state does not account for the fact that subjects with a lower mean response had a higher probability on illness and death, which influence the mean response conditional on being in a state.

8. THE BREAST CANCER DATA EXAMPLE

We return to the data of the breast cancer trial. The transition hazards in the illness–death model of Figure 1 were estimated separately per treatment arm, following the clock forward approach. It was found that the disease-free survival rate and the overall survival rate in the patients of the QOL-health study were slightly better in the high-dose group. The disease-free survival rate at 5 years was 65 per cent in the high-dose arm, compared with 59 per cent in the conventional dose, while the overall survival rates at 5 years were 72 per cent in the high-dose arm and 70 per cent in the standard dose arm.

In the GEE model, treatment group, time and time and treatment interactions were entered in the model. Time was entered as a factor allowing the means in the two treatment groups to vary arbitrarily over time. The weights in the GEE approach were obtained by modeling the probability of non-response π_{ik} , which was done separately for responses in the relapse free and relapse state.

The factors that determined π_{ik} were studied separately per state. It turned out that in the relapse-free state, the probability of response was not related to relapse time or time of death, but patients with lower PH measurements were more inclined to stop responding. Therefore, we assumed that in this state π_{ik} followed a logistic model, with as covariates the previous response $Y_{i(k-1)}$, treatment, time as factor and the interaction between treatment and time. In the relapse state, the probability of missing did not depend on the observed PH values in the relapse-free period and time of relapse, but shortly before death the PH often dropped. In addition, less questionnaires were received shortly before death. For that reason, π_{ik} was modeled after a relapse with a logistic model, with as covariates the previous response $Y_{i(k-1)}$, the disease state at the previous time point $s_{i(k-1)}$, treatment, time (as linear covariate, because of limited number of observations), treatment–time interaction and an indicator variable, equal to 1 if a patient died before the next visit $t_{(k+1)}$ $k = 1, \dots, J - 1$. In this way all data were used, except the last measurement for patients who did not have complete follow-up and were censored in the relapse state ($n = 31$).

There were 35 patients without any PH measurements and some patients had intermediate missing values. The approach of weighted GEE can be extended to non-monotone missings [13, 16], but this is rather complex. Therefore, we imputed intermediate missing values, assuming that the data were multivariate normal distributed, separately per treatment group and disease state. This was done with the *R* package *norm* [17]. For the patients without any PH measurements, no values were imputed; only their data on survival and relapse time were used to estimate the multi-state probabilities.

In Figure 3 the results of the weighted GEE approach are given. The standard errors obtained by bootstrapping (500 replications) were used to calculate the 95 per cent confidence intervals shown in Figure 3. In each bootstrap data set, the multi-state probability and the GEE estimation were recalculated and the intermediate missing values were reimputed.

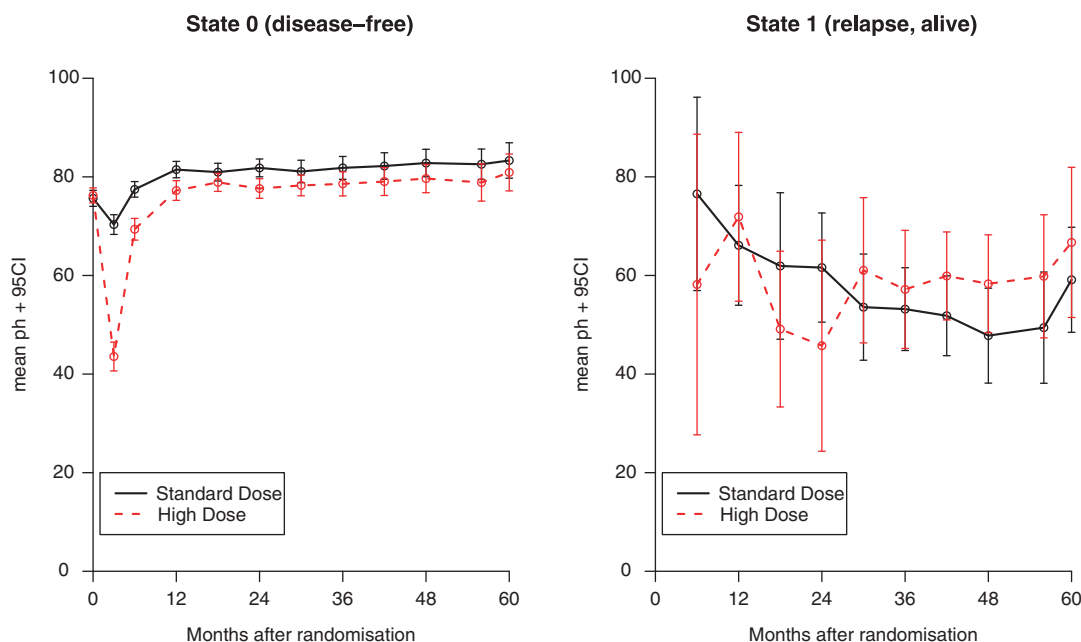


Figure 3. Estimated mean PH, conditional on being in the disease state. The means were estimated using inverse probability weighted GEE.

Figure 3 shows the expected PH for a patient, given the disease state at this time point. Clearly, it is seen in this figure that disease-free patients have a much lower mean PH in the first year (visit 0–3) than patients in the high-dose group. The negative effect of the high dose seems to remain after one year. Note that this is conditional on being relapse free, in the relapse state there is no significant difference between the two treatment groups. Fitting a more restricted model, assuming a constant difference between the two treatment arms after 1 year, we found a statistically significant long-term effect conditionally on being relapse free: the average response with the standard dose was 3.24 (95 per cent CI 0.87; 5.60) higher than with the high dose. The difference after one year conditionally on being in state 1 (alive with relapse) was higher in the high-dose group but not statistically different (difference was 3.16, 95 per cent CI (−4.99; 11.30)).

The estimated marginal means are given in Figure 4. The plots here are population averaged and describe the evolution of the average PH over time in the population of breast cancer patients. Handling measurements after death was done in three different ways: by assigning the value 0 to PH measures after death, by assigning the value 50 and by considering the mean pattern over time, conditional on being alive. Figure 4 shows how the handling of PH after death influences the results. If a value of 0 is assigned to observations after death, the mean PH drops rapidly over time, while conditional on being alive the mean PH after visit 5 hardly changes. In all approaches, however, the difference between the two treatment groups after visit 5 is small. Bootstrapping with 500 resamples was used to obtain standard errors.

To study the joint distribution of disease state and PH, we assumed that PH followed a normal distribution. PH below 75 was defined as low PH. Figure 5 shows the stacked joint probabilities

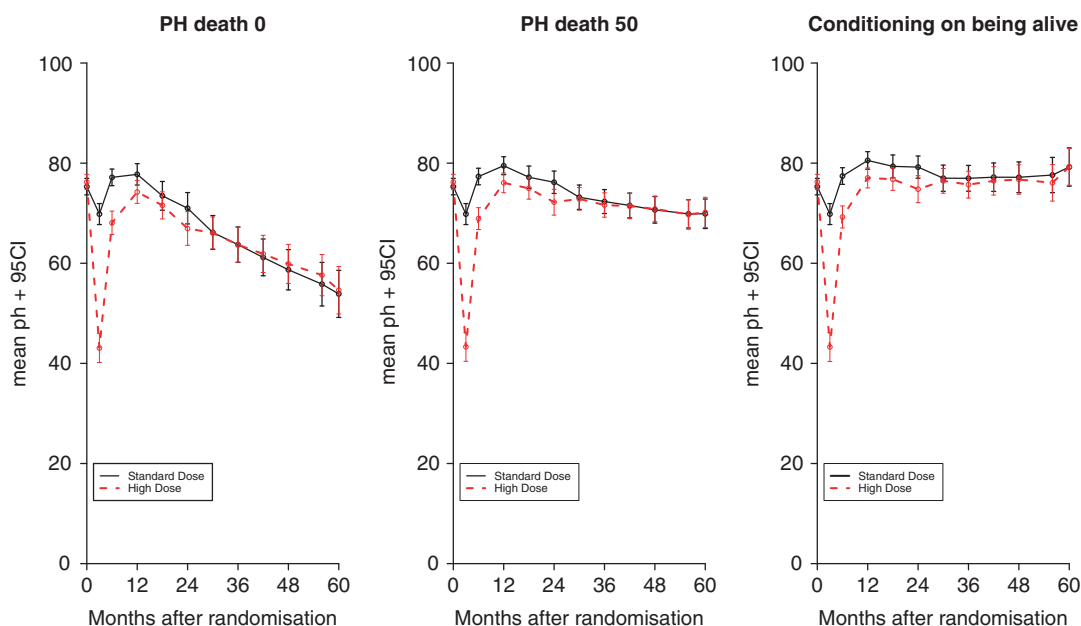


Figure 4. Estimated marginal mean PH, obtained by assigning for PH after death the values 0, 50, or conditional on being alive.

of disease state and PH (high/low) over time in the breast cancer data set and Table I showed the estimated percentages to be in each of the states with low or high PH over time. Here, clearly the large negative impact of high-dose therapy on the PH in the first year is seen. Figure 5 and Table I also show that later during follow-up the percentage of patients who are disease free and with good PH is very similar in both treatment groups (at 5 year, 40 per cent versus 40 per cent, difference 0 per cent, 95 per cent CI -8 percent, $+8$ per cent). At first sight, this seems to contradict the left part of Figure 3, where conditional on being disease free, patients in the standard dose group have a significant higher PH value even after 2 years of follow-up. However, there are relatively more patients with low PH, but disease free in the high treatment group (difference high–low 5 per cent, 95 per cent CI -2 per cent, $+11$ per cent), while there are less patients with low PH and relapse in the high treatment group (difference high–low -3 per cent, 95 per cent CI -7 per cent, 1 per cent). Although these differences are not statistically significant, they could explain the differences between the left part of Figures 3 and 4.

9. DISCUSSION

In this paper we have shown a way of modeling disease state data and longitudinal measurements simultaneously. Death is treated differently from the other disease states. The impact of death on the longitudinal response can be studied either by filling in a predefined value for death, or studying the response conditional on being alive. This is in contrast to using mixed effects models where implicitly values are imputed after death.

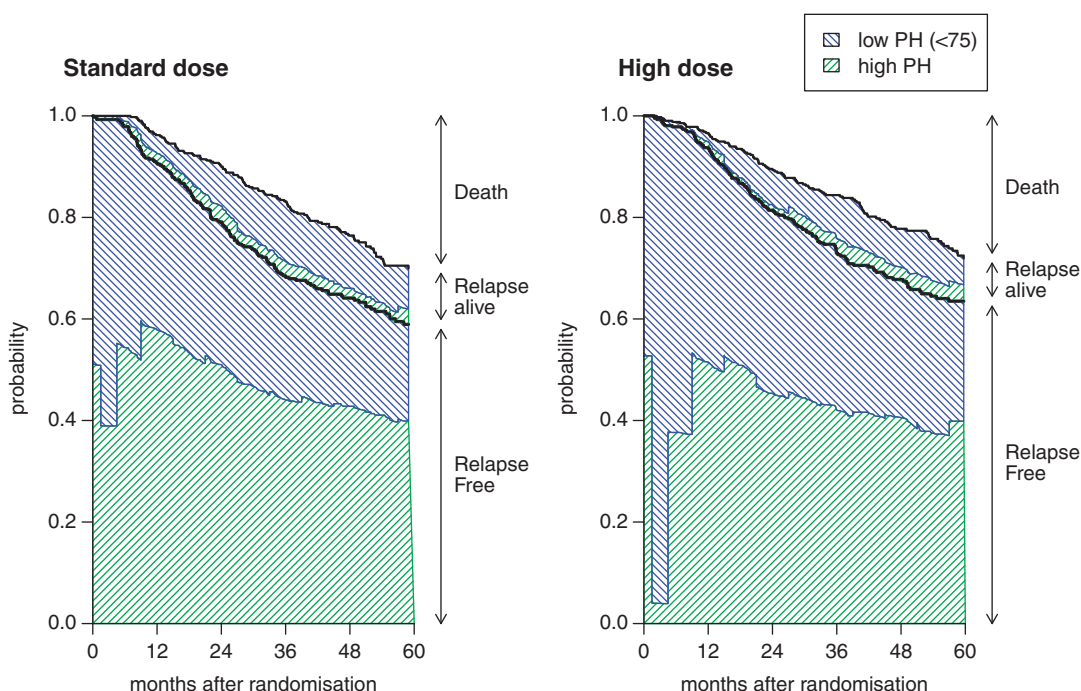


Figure 5. Stacked probabilities to be in each of five different situations: disease free with high PH, disease free with low PH, alive after relapse with high PH, alive after relapse with low PH and death.

Table I. Estimated percentages to be in each of the five different situations over time for the two treatment groups.

	6 months		1 year		2 year		5 years	
	Standard	High	Standard	High	Standard	High	Standard	High
Death	0	1	4	3	10	10	30	28
Relapse, low PH	1	1	3	2	8	7	8	5
Relapse, high PH	1	0	2	1	3	1	3	3
Disease free, low PH	43	60	33	42	28	36	19	24
Disease free, high PH	54	38	58	52	51	45	40	40

Our approach takes into account that the probability of non-response depends on the disease state and that the non-response may depend on different factors in different disease states. A major advantage of our approach is that it clearly shows the effect of the different disease states on the longitudinal response. Furthermore, the marginal effect of a binary exposure variable can be calculated directly and *p*-values and confidence intervals can be obtained by bootstrapping. For continuous covariates the marginal effects are not directly quantified, but the effect of a covariate can be assessed by plotting the marginal means over time for several values of the covariate. An alternative approach to obtain marginal effect parameters could be accounting for disease state and

time of transition into the disease state as auxiliary covariates in the weight model. The comparison of both approaches will be a topic of future research.

We applied weighted GEE models in each separate disease state, and then combine the estimated means over the states. In our example, the number of observations was large, which enabled us to study the long-term differences separately per state. Alternatively, a more parsimonious model could be constructed, assuming the same effect of certain covariates in the different states.

Our analysis depends on several assumptions. We assumed that missing responses due to censored event time are ignorable. This can be assumed in a carefully conducted study where follow-up information on disease events is collected irrespective of the values of the longitudinal measured responses. The events in our example, relapse and death, have a very large impact on patients and are unlikely to be missed even if the patients do not fill in the QOL questionnaires.

Within the states we assumed a missing at random mechanism, which could depend on the exact survival and relapse time. It is not possible to test if these assumptions are valid, but a sensitivity analysis could be performed to see how much the results change if the missing mechanism is not valid. This can be done along the lines of Dufouil *et al.* [5] who included the current measurement with a range of plausible coefficients in the logistic model for non-response.

Relapse and survival times can be censored, which yields problems if the probability of non-response depends on the relapse and survival time. In the breast cancer data set, this was not a major problem, because most patients were followed for the full 5-year period and the probability of non-response in the disease-free state did not depend on the survival time or relapse time. Otherwise, imputation of censored survival or relapse times could be a solution here.

REFERENCES

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
2. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine* 2003; **22**:795–809. DOI: 10.1002/sim.1397.
3. Ribaudo HJ, Thompson SG, Allen-Mersh TG. A joint analysis of quality of life and survival using a random effect selection model. *Statistics in Medicine* 2000; **19**:3237–3250.
4. Billingham LJ, Abrams KR. Simultaneous analysis of quality of life and survival. *Statistical Methods in Medical Research* 2002; **11**:25–48.
5. Dufouil C, Brayne C, Clayton D. Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine* 2004; **23**:2215–2226. DOI: 10.1002/sim.1821.
6. Kurland BF, Heagerty PJ. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* 2005; **6**:241–258. DOI: 10.1093/biostatistics/kxi006.
7. Avlund K, Vass M, Kvist K, Hendriksen C, Keiding N. Educational intervention towards preventive home visitors reduces functional decline in community-living older women. *Journal of Clinical Epidemiology* 2007; **60**:954–962. DOI: 10.1016/j.jclinepi.2005.06.015.
8. Diggle P, Farewel D, Henderson R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Applied Statistics* 2007; **56**:499–550. DOI: 10.1111/j.1467-9876.2007.00590.x.
9. Rodenhuis S, Bontenbal M, Beex LVAM, Wagstaff J, Richel DJ, Nooij MA, Voest EE, Hupperets P, van Tinteren H, Peterse HL, TenVergert EM, de Vries EGE. High-dose chemotherapy with hematopoietic stem cell rescue for high-risk breast cancer. *New England Journal of Medicine* 2003; **349**:7–16.
10. Buijs C, Rodenhuis S, Seynaeve CM, van Hoesel QG, van der Wall E, Smit WJ, Nooij MA, Voest E, Hupperets P, TenVergert EM, van Tinteren H, Willemse PH, Mourits MJ, Aaronson NK, Post WJ, de Vries EG. Prospective study of long-term impact of adjuvant high-dose and conventional-dose chemotherapy on health-related quality of life. *Journal of Clinical Oncology* 2007; **25**:5403–5409. DOI: 10.1200/JCO.2007.11.2813.
11. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; **26**:2389–2430. DOI: 10.1002/sim.2712.
12. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **7**:13–22.

13. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
14. Sherman M, le Cessie S. A bootstrap for correlated outcomes in generalized linear models. *Communication in Statistics, Simulations and Computations* 1997; **26**:901–926.
15. Glasziou JP, Cole BF, Gelber RD, Hilden J, Simes RJ. Quality adjusted survival analysis with repeated quality of life measures. *Statistics in Medicine* 1998; **17**:1215–1229.
16. Shardell M, Miller RR. Weighted estimating equations for longitudinal studies with death and non-monotone missing time-dependent covariates and outcomes. *Statistics in Medicine* 2008; **27**:1008–1025. DOI: 10.1002/sim.2964.
17. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, 1997.